PERCEPTUALLY OPTIMISED SIGN LANGUAGE VIDEO CODING

D. Agrafiotis, N. Canagarajah, D.R. Bull

Image Communications Group, CCR, University of Bristol, Bristol, BS8 1UB, UK

ABSTRACT

Mobile video telephony will enable deaf people to communicate in their own language, sign language. At low bit rates coding of sign language video is challenging due the high levels of motion and the need to maintain good image quality to aid with understanding. This paper presents perceptually optimised coding of sign language video at low bit rates. The proposed optimisations are based on an eye-tracking study that we have conducted with the aim of characterising the visual attention of sign language viewers. Analysis and results of this study and two coding methods, one using MPEG-4 video objects and the second using foveation filtering, are presented. Results with foveation filtering are promising, offering a considerable decrease in bit rate in a manner compatible with the visual attention patterns of deaf people, as these were recorded in the eye tracking study.

1. INTRODUCTION

Third generation mobile networks will enable wireless transmission of compressed video data. This will be of great benefit to deaf people as it will allow them for the first time to communicate anytime / anywhere in their own language, sign language. Sign languages are visual languages. As such they demand good image quality to aid with understanding. Many video coding systems have focused on the compression of typical video conferencing sequences where a head and shoulders view of the participant is usually involved. Sign language image sequences include, in addition, the rapidly moving hands and arms of the imaged signer resulting in increased bit rate requirements [1]. Bandwidth however in 3G networks might be (at least initially) limited to as little as 64 Kbits/sec.

Video coding at low bit rates will always result in some information being lost in order to satisfy rate requirements set by the network over which transmission will take place. However, it may be possible to localise information loss based on perceptual criteria, i.e. allow information to be lost where it does not significantly impair signed language comprehension, and keep information where it is essential for comprehension.

This paper presents results for perceptually optimised coding of sign language video at low bit rates. The aim of the proposed optimisations is to increase the comprehension potential of the coded sequences at such low rates. In order to optimise coding we have conducted eye-tracking experiments to study the visual attention of sign language viewers and thus obtain some information about the visual importance of various regions in the video frame. In this paper we present results from these experiments and describe an optimised region-based coding approach based on these results. The structure of this paper is as follows: first the eye-tracking study is described, the results of the study are analysed, and their relevance to video coding is established. Then two region based coding approaches are described, one using the video objects (VOs) aspect of MPEG-4 and the second using foveated filtering prior to coding. Results with both approaches are presented. Finally conclusions are given and further work is suggested at the end of this paper.

2. EYE TRACKING STUDY

Eleven subjects from 4 different categories - 3 Deaf from Deaf parents (DD), 3 Deaf from Hearing parents (DH), 3 Hearing Signers (HS), and 2 Hearing Beginners in sign language (HB) - took part in the experiments, which involved watching 4 short narratives in British Sign Language (BSL). The video material was shot in a studio with the signer sitting in front of a blue screen and wearing plain clothes. The duration of each clip was around 1 minute. The clips were captured in the Digital-S format and were subsequently converted and displayed in the CIF format - 352x288, 4:2:0, uncompressed.

The Eyelink system [2] was used to record the participants' eye-gaze while watching the 4 video clips. The system briefly consists of a headband with two miniature high speed cameras which provide binocular eye-tracking and a third one which tracks 4 IR markers mounted on the stimulus display and enables motion compensation and true gaze position tracking. An operator PC with a DSP card analyses the captured images at a sampling rate of 250Hz. Software written for the experiments reports eye-gaze data for each frame of the video clip (10 samples per frame with a frame rate of 25fps). The experiments took place in a darkened room with the participants sitting in front of 21" screen (39x29cm active visible area) with the resolution set at 640x480, at a distance of 60 cm away from it. A chinrest was used in order to ensure the best possible gaze position accuracy.

0-7803-8163-7/03/$17.00 © 2003 IEEE

2.1. Results

Analysis of the results, and mainly of the fixation location and duration (i.e. locus at which eye-gaze is directed) show that sign language viewers, excluding the hearing beginners, seem to concentrate on the facial area and especially the mouth. In fact participants from the DD and HS categories never looked at the hands while those from the DH category showed just a small tendency to look at the hands. In contrast hearing beginners did look at the hands more frequently (mainly due to a lack of understanding). Figure 1 shows the vertical position of the recorded fixation points of subjects from different categories relative to the position of the video on the screen - i.e. y = 0 corresponds to the first row of pixels of the video clip – for the first 250 frames of clips 2 and 4. Points with y > 150 are fixation points below the face and mostly on the hands. It can be seen that only a few points cross this boundary, and only a few subjects – in particular subject 1 of the DH category – actually fixate on the hands during the sequence (2 or more points > 150). The eye tracking results are in accordance with anecdotal reports coming from deaf people which indicate that signers maintain their visual attention on the face of the person signing.

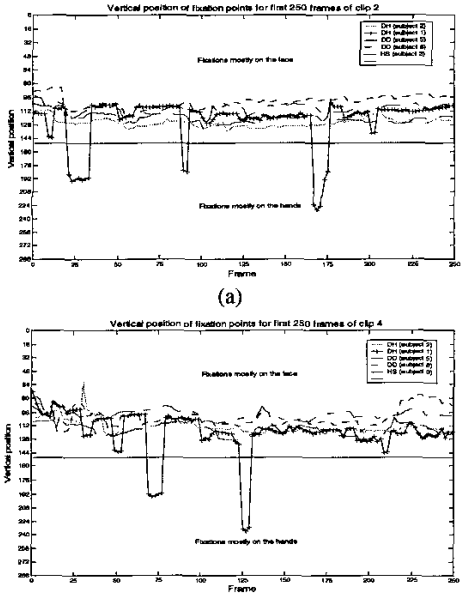


Figure 1. Vertical position of fixation points relative to location of video for the first 250 frames of clips 2 (a) and 4 (b).

3. OPTIMISED CODING

Guided by the above eye tracking results we decided to follow a region based coding approach whereby different quality is assigned to the different regions of the video frame The fact that the centre of the visual attention of sign language viewers is located on the face suggests that greater "coding attention" should be given to it relative to the rest of the video frame. In other words, for a fixed bit budget more bits should be spent on coding the face (especially the area around the mouth) and less for the rest of the image, namely the body and hands of the signer and the background. Using MPEG-4 as the base codec two ways of assigning variable priority to the different regions of the image have been explored. The first method exploits the object-based capability of MPEG-4, while the second applies a variable low pass filtering to the data prior to coding.

3.1. MPEG-4 and Video Objects

Two test sequences were captured with the signer sitting in front of a plain blue background (figure 2). Both sequences were stored at CIF resolution with 25 fps. In one of the two sequences the signer was wearing gloves in order to facilitate accurate segmentation of the hands. Both sequences were subsequently segmented into face, body and hands, (constituting the foreground) and background video objects.

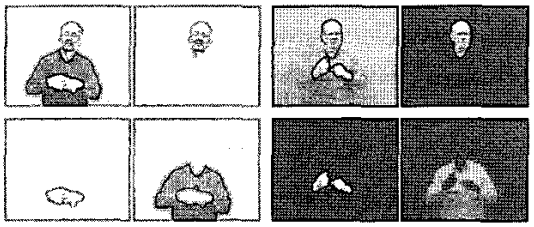


Figure 2. Test sequences and segmented video object planes. Sign1 sequence (left), Sign2 sequence (right).

Initially, MPEG-4 frame-based coding of a CIF resolution sign language sequence with a plain-background is investigated. The first frame is coded in Intra mode with the rest being inter coded. Motion compensation is done with 1/4 pixel accuracy and a search window of 16 pixels. The output frame rate is 12.5 fps. When coded at around 64 kbps with quantisation parameter (QP) of 30, the video was judged to be of poor quality, mostly due to distortion around the face. The sequence was then segmented into foreground and background. The foreground only was then coded as a VO with MPEG-4, which uses lossless shape coding, shape adaptive DCT and padding for the borders. The overhead due to shape coding was found to be around 13 kbps. Comparing the foreground only PSNR of the frame-based coded and the VO coded video, an 8 kbps overhead was observed for the VO version with the offset coming from better texture coding at the edges (Figure 3). At high overall bit-rates, the compression performance becomes almost identical to frame-based coding, since the shape coding overhead is constant. In sequences with background changes, the coding of the foreground as a separate VO will result in greater savings since background changes are not coded. Further segmentation of the foreground VO into hands, face, and body was then carried out. The overhead for shape coding in this case is

approximately 3.2, 7.5 and 17.9 kbps for the Sign1 face, hands and body VOs respectively, giving a total of 28.6 kbps for shape information. For Sign2, the shape coding overheads are 4.2, 4.4 and 16.4 kbps, respectively, resulting in a total of 25 kbps. A similar as before comparison (with the three VOs comprising the foreground, and all coded with the same QP) showed a 10kbps overhead for the VO coded Sign2 sequence (Figure 3). In order to assign a higher priority to the face, a lower quantiser should be used with the face VOP while a higher one is used for the hands and body VOPs (lower priority). As before the background can be either discarded or assigned a very small proportion of the available bit budget. The disadvantage of having an additional overhead with VOs is counterbalanced by this ability to assign a different quantisation parameter, and hence quality, to each of them. However the problem of real time segmentation with varying scene content is still a major obstacle in any attempts to apply VO coding.

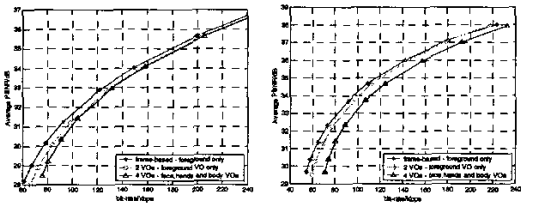


Figure 3. Foreground only PSNR of the frame-based, 2-VO and the 4-VO coded video. Left Sign1, right Sign2.

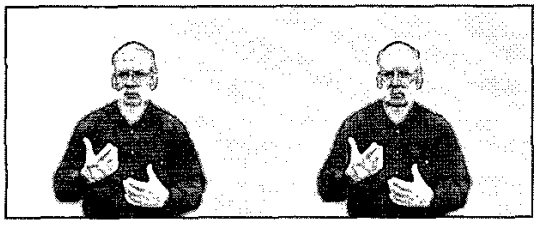


Figure 4. Decoded frame of 4-VO coded Sign1 sequence. Left, all VOs have the same quantiser of QP=10. The total bit-rate is 127.60 kbps, with the face at 24.1, hands at 53.1 and body at 54.2 kbps. Right, VOs have different quantisers. Total bit-rate is 123.70 kbps, with the face at 42.86 (QP=6), hands at 40.29 (QP=14) and body at 40.55 kbps (QP=30).

3.2. Foveated Pre-filtering

Recently there has been growing interest in foveated video compression, where the fall off in spatial resolution of the human visual system away from the point of fixation is exploited for the reduction of the bandwidth requirements of compressed video [3][4]. By removing spatial high frequency components from regions away from the centre of fixation the entropy of the video (after quantisation) is lowered, allowing increased compression gain. The foveation regions of an image (regions of constant perceived resolution) are determined using a contrast threshold formula [5], which is based upon fitting experimental human contrast sensitivity graphs measured as a function of spatial frequency and retinal eccentricity:

$$CT(f,e) = CT_0 \exp\left(af\frac{e+e_2}{e_2}\right) \qquad (1)$$

f is spatial frequency (cycles per degree), e is the retinal eccentricity (degrees), CT_0 is a constant minimal contrast threshold, a is a spatial frequency decay constant, e_2 is the half-resolution eccentricity and $CT(f,e)$ is the visible contrast threshold as a function of f and e (that is the minimum contrast required for spatial frequency f to be visible at eccentricity e). The best suggested fitting parameters [5] are: a = 0.106, e_2 = 2.3, and CT_0 = 1/64. Contrast sensitivity $CS(f, e)$ is the inverse of the contrast threshold. By setting the contrast threshold $CT(f, e)$ to 1 (maximum contrast), equation (1) can be used to find the critical (cut-off) frequency fc for a given eccentricity e, i.e. the maximum detectable frequency at this eccentricity.

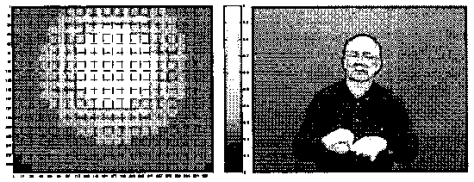
$$f_c = \frac{e_2 \ln(1/CT_0)}{(e+e_2)a} \quad (cycles/degrees) \qquad (2)$$

Use of the above equation leads to the local bandwidth approach to foveating an image. We have followed the local bandwidth method described in the work of Sheikh et.al [4]. This method involves pre-filtering of the video frame before passing it to the encoder. The image is partitioned in different filtering regions based on their distance from the centre of fixation. Filtering is then applied to each region (except for the fixation region) with the cut-off frequency getting lower for these regions that lie further away from the fixation point. The cut-off frequency of each region is given by :

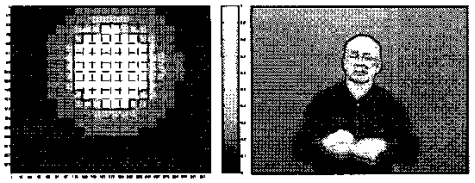
$$f_c = \frac{e_2 \ln(1/CT_0)}{\left(\tan^{-1}\left(\frac{d(x)}{Nv}\right)+e_2\right)a} \qquad (3)$$

where $d(x)$ is the Euclidean distance of pixel x from the fixation point, N is the width of the image, v is the viewing distance (in 100s of pixels) and $\tan\text{-}1(d(x)/Nv)$ is the eccentricity e. All distance and coordinate measurements are normalised to the physical dimensions of pixels on the viewing screen. For a number of supported viewing distances v, 8 possible values of the maximum detectable frequency are used, effectively partitioning the image into a set of foveated regions (maximum 8) such that each region has a constant maximum detectable frequency. For computational speed the foveation regions are constrained to be the union of disjoint macroblocks (16x16 pixel blocks), and the boundary (the distance from the fixation point) of each foveation region is pre-calculated for all the possible values of the viewing distance v. Using this model, the image is essentially filtered with variable filters of a different cut-off frequency. Similar filters with [4] have been used. The normalised cut-off frequency of each filter is equal to i/8 where i is the region index (from 1/8 to 7/8). Symmetric extension is used for filtering the image at the edges. Filtered samples at the boundaries of foveated regions are averaged in order to avoid the creation of sharp edges. In theory the cut-off frequency

should match the maximum detectable frequency of the HVS which is dependent on the viewing distance v. In practice however we control the amount of filtering (and hence the introduced distortion) by altering the value of v and CT_0 (figure 5). The unfiltered region radius is 64 pixels and the estimated fixation point lies on the marker positioned above the mouth of the signer.



$v = 3 - CT_0 = 1/64$



$v = 1 - CT_0 = 1/64$

Figure 5. Controlling foveation through viewing distance v and constant threshold CT_0.

In order to evaluate the effect of variable pre-filtering on sign language video compression three test sequences were used. One of them was Sign1 and the other two were generated from Sign1 through the inclusion of different backgrounds, one static and one moving. Figure 6 shows the bit rate of each coded sequence for specific values of the quantisation parameter with and without foveation.

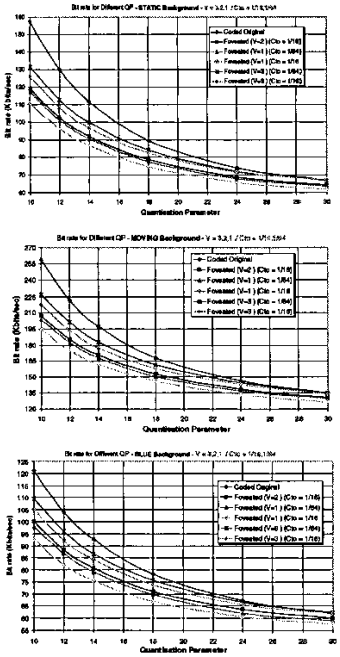


Figure 6. Bit rate versus quantisation parameter graphs for normal and foveated frame based coding.

As the results indicate, the reduction in the bit rate can be as high as 29% (static background, v=1,CT0 = 1/16, QP 10). The bit rate reduction depends on the amount of filtering applied (amount of foveation) and the quantisation parameter used. For high values of QP, foveation can only reduce the bit rate by a maximum of 7%, since the quantiser step size is large and many of the higher frequency coefficients of the DCT transform become zero any way. For a smaller quantisation parameter, even a very small amount of foveation reduces the bit rate by about 10 to 15% (e.g. static background, QP 10, v=3, CT0 = 1/64 : 131.54 kbps vs. 157.33 kbps for the original coded sequence). One of the advantages of this method is that it removes the need for segmentation. Instead knowing that the centre of fixation lies on the facial area, we can roughly track the face and use a point on it as the estimated fixation point.

4. CONCLUSIONS

In this paper we proposed optimised coding approaches for sign language video compression using MPEG-4 as the base codec. The suggested optimisations are based on the results of the eye tracking study that we have conducted, which showed that sign language viewers concentrate mostly on the face and especially the mouth. Coding results using video objects show that coding of sufficient quality can be achieved at rates close to 128 Kbps for a CIF resolution video, without coding the background. However video objects incur an overhead due to shape coding, which for CIF resolution sign language sequences is around 10 kbps. When video objects are used, a lower quantiser (higher priority) should be assigned to the face. In order to avoid the need for segmentation imposed by a video object coding approach, foveated frame based coding was suggested. Foveated coding requires only rough tracking of the signer's face for estimation of the fixation point. The amount of foveation/filtering can be controlled in a scalable manner through the value of two parameters. Results showed that with even a small amount of foveation a 15% reduction in bit rate can be achieved.

5. REFERENCES

[1] R.P. Shumeyer, E.A. Heredia, K.E.Barner, "Region of Interest Priority Coding for Sign Language Videoconferencing", IEEE Wshp. on M/media Signal Processing, pp. 531-536, 1997.

[2] Sensor Motoric Instruments (SMI), "Eyelink System User Documentation" ,1999.

[3] W.S.Geisler, J.S.Perry, "Variable-Resolution Displays for Visual Communication and Simulation", Society for Information Display, vol. 30, pp. 420-423, 1999.

[4] H.R. Sheikh, S.Liu, B.L.Evans, A.C.Bovic, "Real Time Foveation Techniques for H.263 Video Encoding in Software", IEEE ICASSP, vol.3, pp. 1781-1784, 2001.

[5] Wilson S. Geisler, Jeffrey S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication", SPIE Proceedings, vol. 3299, 1998.